8. Smith BH. Cult-lit: Hirsch, literacy, and the "national culture." *South Atl Q* 1990;89:69–88.
9. Postman N. Cultural literacy. *The Atlantic* 1989;264:119–124.
10. Fleming DB. Beware of those bearing lists. *N ASSP Bull* 1992;76: 105–109.
11. Kohl H. Rotten to the Core. *The Nation* 1992;254:457–462.
12. Shenker J. The well-stocked brain (book reviews). *NY Times (Print)* 17 December 1989;Section 7:22, column 1.
13. Seligman D. Getting it wrong. *Fortune* 1993;128:169–170.
14. Klinghoffer D. Trivial pursuit: the pros and cons of cultural literacy. *Natl Rev* 1999;51:61.
15. Bloom A. *The Closing of the American Mind: How Higher Education Has Failed Democracy and Impoverished the Souls of Today's Students*. New York: Simon and Schuster, 1987.
16. Urban WJ, reviewer. *J Am Hist* 1988;75:869–874. Review of: Bloom A. *The Closing of the American Mind: How Higher Education Has Failed Democracy and Impoverished the Souls of Today's Students*; Hirsch ED Jr. *Cultural Literacy. What Every American Needs to Know*; Ravitch D, Finn CE Jr. *What Do Our 17-Year-Olds Know? A Report on the First National Assessment of History and Literature*.
17. Seaton J. Cultural conservatism, political radicalism. *J Am Cult* 1989;12:1–9.
18. Pentony JF. Validity and reliability of the Cultural Literacy Test. *Psychol Rep* 1996;78:1027–1033.
19. Bart WM. On the relationship between cultural literacy and scholastic achievement among university students in secondary education. *Psychol Rep* 1998;82:841–843.

Validation Studies: Bias, Efficiency, and Exposure Assessment

Nilanjan Chatterjee and Sholom Wacholder

Measurement error is the bane of epidemiologic studies of diet, behavior, and environmental factors. Even molecular assays—whether for simple genotypes or for complicated biochemistry—are not immune. In this issue, Stürmer *et al.*[1] make some useful observations on methods to reduce or eliminate the bias from measurement error. Herein, we discuss the general problem of measurement error and comment on the current state of epidemiologic methods to mitigate its effect.

Errors in variables (the term used by statisticians) lead to distorted estimates of effect and to underpowered or biased tests. The impact of errors is well understood, at least regarding the direction of the estimate, in a few important special situations. However, measurement error can introduce unpredictable distortions in many realistic settings.[2,3]

Case-control studies with exposure data or biospecimens collected retrospectively are particularly vulnerable to poor measurement of exposure. *Nondifferential* measurement error is often a consequence of exposure information that is collected long after the exposure occurs. *Differential* error can arise when symptoms or treatment of disease affect a biomarker (other than germ-line DNA) or, along with knowledge of the presence of disease, influence response to questionnaires.

Internal validation studies can be used to reduce the impact of measurement error. An error-prone exposure measurement Z is collected from everyone in the main study. A more accurate but more expensive measurement X is also available, in principle, for everybody. However, owing to cost or practical considerations, X is collected only on a *validation sample*, consisting of small subsets of cases and controls selected randomly. Clearly, the risk parameter associated with X could be estimated unbiasedly but quite imprecisely with a *complete case estimator* (CCE) that discards all the imperfect but informative Z measurements from subjects not in the validation sample. The *regression calibration* (RCE) and *semiparametric efficient estimators* (SPE) exploit the imperfect measurements Z from individuals who were not included in the validation sample to obtain a more efficient estimate of the risk parameter.

What is the basic principle behind these more sophisticated "bias correction" methods that use both sets of measurements? The validation sample reveals the relation between Z and X. Based on this relation, a probabilistic distribution of X can be inferred from Z for subjects with unknown X. SPE and RCE use different ways of predicting X from Z; they make different tradeoffs between stronger assumptions about the structure of the error and greater reliance on the validation data itself. RCE also requires an additional assumption that the measurement error be small.

From the Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

Address correspondence to: Nilanjan Chatterjee, 6120 Executive Blvd, EPS 8038, Bethesda, MD 20892-7244; chattern@mail.nih.gov

DOI: 10.1097/01.EDE.0000022948.80077.AE

Understanding the basis of the nomenclature can help one understand the distinction. The term regression calibration describes how Z is calibrated to X based on the parametric regression model for X given Z; incidentally, it also evokes the calibration of the regression of interest by data from the validation study. In contrast, SPE predicts the distribution of X given Z from the validation study *nonparametrically*, that is, without imposing a parametric relation between X and Z. It is deemed *semiparametric* because it involves one parametric and one nonparametric component: the parametric component is the regression model for Y given X, and the nonparametric component is the distribution of X given Z. The method is called *efficient* because it is defined as the most efficient among the class of all semiparametric estimators that treat the distribution of X given Z nonparametrically. In other words, SPE predicts the distribution of X given Z from the validation study nonparametrically, whereas RCE requires a specific parametric assumption about the conditional distribution of X given Z, such as the conditional mean of X being linear in Z, or what Stürmer *et al.* call *linear measurement error*.[1]

An analogy can be drawn with the use of a parametric *t*-test or a nonparametric Wilcoxon rank-sum test when testing whether the distributions of a variable are the same in two groups. The *t*-test relies on the assumptions that the variables are normally distributed with equal variance in both groups; it is more efficient when the assumptions hold, but can be quite misleading when they are strongly violated. In contrast, the nonparametric procedure makes no distributional assumption; its superior performance when the *t*-test assumptions are violated compensates for its lower efficiency when the assumptions hold. In bias correction, the precision of RCE under the correctly specified error model contrasts with the robustness of SPE against departures from the model. The nonparametric aspect of SPE allows it to capture more nuances of the relation between X and Z than RCE. Assume, for example, that multidimensional X is observed with nondifferential measurement error in each component and that Z is the corresponding error-prone variable. Even if X and the errors Z–X are not normally distributed and each of the components of X and Z–X are correlated with each other,[2,3] semiparametric estimators remain valid. When the normality assumption is correct and the relation between Z and X is linear, however, SPE is less efficient than RCE because RCE capitalizes on the imposed structure. Stürmer's simulation studies (Table 2) clearly show that for small measurement error, the efficiency gain of RCE over SPE can be substantial if the assumed linear measurement error model holds, but the authors do not explore the consequence of the violation of this assumption.[1]

Differential measurement error poses an additional challenge. SPE, although far superior to RCE in the typical nonparametric setting, may not be fully robust, as it requires the investigator to specify the form of differential measurement error. Stürmer *et al.*, for example, considered a normal measurement error distribution where the difference in distributions between cases and controls is characterized either by a shift in the mean or a change in the scale of the underlying normal distribution.[1] Although Stürmer's simulation experiments show that SPE is unbiased under this form of differential measurement error,[1] SPE may not be robust when the errors are not normally distributed. The complete case estimator is fully robust to differential measurement error, as it relies only on the gold-standard measurements.

More often than not, X is an alloyed-gold standard,[4] an imperfect approximation of a gold standard. In this situation, the "bias correction" procedures estimate the same regression that they would if X were used for everyone in the study. Therefore, the measurement error in X leads to residual bias in the estimate. When the alloyed gold standard is an "unbiased" predictor of the true gold-standard, both the CCE and SPE estimators are attenuated in Stürmer's simulations (Table 8).[1] Here, *unbiased* means that the average value of repeated measurements of X will give the true gold standard for each individual. This no-bias assumption, however, can easily be violated in practice. For example, error correction based on use of diaries for sun exposure or diet can modify behavior through a kind of personal "Hawthorne effect."[5] The task of scrupulously recording food intake may in itself affect eating behavior during the diary period so that the recorded information is unrepresentative of usual diet beforehand or afterwards.

"Many epidemiologists regard validation studies principally as a way to correct bias from measurement error. In contrast,. . .[we] view validation designs as tools for minimizing the cost of a study with fixed power. . ."

Many epidemiologists regard validation designs principally as a way to correct bias from measurement error.

In contrast, we usually take for granted the ability of these methods to produce an unbiased estimate, at least under idealized conditions. We view validation designs as tools for minimizing the cost of a study with fixed power or, equivalently, for maximizing the precision of the estimate of the main study parameter with fixed cost. We therefore consider it valuable to consider designs with stratified random sampling (with strata defined jointly by case-control status and error-prone exposure measurements), which can often be substantially more efficient than either simple random sampling or standard case-control sampling.[6–8] For example, oversampling cases so that the numbers of cases and control in the validation stage are equal is clearly a more efficient strategy for CCE, and, not surprisingly, for RCE and SPE as well, than the simple random sampling Stürmer *et al.* considered. Sampling based on Z could further improve the efficiency of the design. For example, if Z is at least moderately correlated with X, oversampling extreme values of Z yields greater numbers of extreme Xs and, therefore, will be more informative for estimating a slope. There need not be any bias attributable to stratified sampling if appropriate statistical methods are used in the analysis stage.

The validation design is a special case of "two-phase stratified study designs," which can provide cost savings in many epidemiologic studies. The two phases are the collection of a set of inexpensive covariates Z for all subjects, followed by the collection of more expensive covariates X at phase 2 on a smaller subsample of subjects selected based on values of Z and case-control status. More generally, two-phase designs can be used profitably to collect information on an expensive exposure of interest,[9] confounder,[6] or effect modifier[10] on a sample of subjects, with the sampling faction varying according to the value of variables available for everyone in the study. Even "old-fashioned" matching and contemporary countermatching[11] can be seen as two-phase strategies,[9] because collection of exposure X depends on the matching variables Z. In the class of two-phase designs, the validation study is special in only one rather trivial way: Z is not included in the risk model, because it is assumed that there is no information about risk of disease attributable to Z that is not contained in X.

So where are we now? The paper by Stürmer *et al.* shows the promise of sophisticated statistical methods for error correction.[1] In general, semiparametric estimators are more flexible and robust than RCE in the presence of poorly understood error mechanisms. When computation of a semiparametric efficient estimator is overly complex, slightly less efficient but simpler semiparametric estimators based on pseudolikelihood methods[12] can be attractive. Software is now available in S-PLUS (MathSoft, Inc., Seattle)[13] for various semiparametric methods of the general two-phase data problem (including validation studies for nondifferential measurement error problems) using the logistic regression model. Further research is needed to establish the robustness of the procedures in realistic settings, specifically for the differential measurement error and the alloyed gold-standard problems, and for determining optimal designs for selecting a validation sample.

We believe that these statistical methods for "bias correction" are ready to be used in case-control studies in some limited situations. In particular, RCE can be an efficient tool when measurement error is small and the error structure is reasonably well understood. If the measurement error is large or the error structure is not known—as often is the case in practice—a semiparametric estimator can be used as a robust alternative, at least when there is no important differential measurement error. For additional economy, an efficient design using stratified sampling should be considered as a way to select the most informative subjects in the validation sample. Evaluation of the performance and utility of these procedures will require several applications in studies with hard-nosed critiques of the validity of the underlying assumptions.

Validation studies, whether or not designed for bias correction, can be crucial to increasing the value of epidemiologic studies. An instructional example is the rapid progress in understanding the epidemiology of cervical neoplasia that followed the identification of polymerase chain reaction as a sensitive and specific assay for infection with oncogenic human papillomavirus[14] through intra- and interlaboratory studies of replicability.[15] The best way to reduce bias from measurement error is to improve tools for measuring exposures including biological markers, environmental samples, and questionnaires.

References

1. Sturmer T, Thurigen D, Spiegelman D, Blettner M, Brenner H. The performance of measurement error correction methods for the analysis of case-control studies with validation data; a simulation study. *Epidemiology* 2002;13:507–516.
2. Kipnis V, Freedman LS, Brown CC, Hartman AM, Schatzkin A, Wacholder S. Effect of measurement error on energy-adjustment models in nutritional epidemiology. *Am J Epidemiol* 1997;146:842–855.
3. Wacholder S. When measurement errors correlate with truth: surprising effects of nondifferential misclassification. *Epidemiology* 1995;6:157–161.
4. Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. *Am J Epidemiol* 1993;137:1251–1258.
5. Neale RE, Green AC. Measuring behavioral interventions by questionnaires and prospective diaries: an example of sunscreen use. *Epidemiology* 2002;13:224–227.
6. White JE. A two-stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol* 1982;115:119–128.
7. Breslow NE, Cain C. Logistic regression for two-stage case-control data. *Biometrika* 1988;75:11–20.
8. Holcroft CA, Spiegelman D. Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified. *Biometrics* 1999;55:1193–1201.

9. Weinberg CR, Sandler DP. Randomized recruitment in case-control studies. *Am J Epidemiol* 1991;134:421–432.
10. Wacholder S, Weinberg CR. Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. *Biometrics* 1994;50:350–357.
11. Langholz B, Borgan O. Counter matching: a stratified nested case-control sampling method. *Biometrika* 1995;45:69–79.
12. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Appl Stat* 2002;48:457–461.
13. *S-Plus* [computer program]. Seattle: MathSoft, Inc., 1996.
14. Schatzkin A, Freedman L, Schiffman M. An epidemiologic perspective on biomarkers. *J Intern Med* 1993;233:75–79.
15. Schiffman M, Herrero R, Hildesheim A, *et al*. HPV DNA testing in cervical cancer screening: results from women in a high-risk province of Costa Rica. *JAMA* 2000;283:87–93.